ChinaXiv合作期刊
图书情报工作
第 65 卷 第 9 期 2021 年 5 月

古籍同事异文的自动发掘研究*

■ 梁媛[1,2] 王东波[1,2] 黄水清[1,2]
[1]南京农业大学信息管理学院 南京 210095 [2]南京农业大学人文与社会计算研究中心 南京 210095

摘 要：[目的/意义]异文是古籍中的常见现象，也是重要研究对象。传统的古籍校勘是从大量古籍文献中人工查找校勘资料包括异文等，不仅耗时、费力、工作量大，而且找到的数据未必精准全面。通过计算机实现异文的自动发掘，可以从更大规模的语料中获取有效信息。并且，结合异文自动发掘的校勘方式可以实现穷尽式检索，对于古籍他校法具有重要意义，为新时期古籍校勘研究提供了新思路和新方法。[方法/过程]本研究以《春秋》及“春秋三传”作为实验语料，引入常用于文本翻译领域的平行语料库思想，结合深度学习算法，对 LSTM、BERT 模型与较为经典的 SVM 模型进行比较实验，并对两部古籍中用不同表述描述同一事件的同事异文相关内容展开进一步探索和讨论。[结果/结论]实验得到适用于“春秋三传”的同事异文自动发掘深度学习模型，证明深度学习等新兴技术融合到古籍知识库构建等研究中的可行性，同时，深度学习技术和平行语料库思想的结合在异文研究中能够发挥较大作用，对数字人文在汉语言文学研究中的应用提供实践支撑。

关键词：春秋三传 异文 BERT 自动发掘 数字人文
分类号：G255.1
DOI：10.13266/j.issn.0252-3116.2021.09.011

1 引言

我国古籍资源丰富，著作不胜枚举，异文现象十分常见。广义上，无论是同一文献不同版本的用字差异，还是同一事件被引用、叙述时遣词用句的区别都可以包含在异文之中[1]。在文字学意义上，由异文可明通假字、古今字和异体字；在词汇学上，由异文可明同义词、同源词和连绵词；在语法学上，由异文可明词序和某些特殊的句法关系。不管是文字学、词汇学还是语法学，都可能会运用到异文的相关研究成果[2-3]。

作为中国古代的主流意识思想，儒家思想对中国甚至世界都有着深远的影响。而《春秋》是儒家最为经典的著作之一，也是我国最早的编年体史书，记录着传统文化和古人智慧。“春秋笔法，微言大义”，为《春秋》作传的“春秋三传”则将这些晦涩难懂的内容补充得全面生动、恰到好处，从历史背景、社会风貌和政治礼制等方面为读者全方位展现了一个多彩春秋。

本研究从《春秋》及“春秋三传”入手，主要针对两部古籍之间对同一事件的不同表述，即同事异文进行研究，并引入常用于文本翻译领域的平行语料库的思想，结合深度学习算法，对古籍中的同事异文自动发掘等相关内容展开进一步探索，以期将新技术融合到古籍知识库构建等研究中，为其提供新的思路和方法。

2 相关研究

异文的研究最早可以追溯到古人对于古籍进行的注疏，到了近现代，很多学科的研究都会涉及与异文相关的内容。正所谓“知其然，知其所以然”，异文的成因一直是相关研究者较为重视的研究方向，罗积勇[4]总结了古籍中异文的 4 种成因，提出了异文不但对校勘和训诂具有重要价值，更是古文释义的重要参考资料。在这之后，王彦坤[5]和石云孙[6]分别针对古书异文和近现代话语中的异文通假现象从不同角度展开研究，并对异文成因等内容进行了分析。

古籍种类繁多，经、史、子、集四部又下分四十四类，将不同类别古籍中的异文作为研究对象，可能会产

* 本文系国家社会科学基金重大项目“基于《汉学引得丛刊》的典籍知识库构建及人文计算研究”（项目编号：15ZDB127）和国家自然科学基金面上项目“基于典籍引得的句法级汉英平行语料库构建及人文计算研究”（项目编号：71673143）研究成果之一。
作者简介：梁媛（ORCID：0000-0002-4922-6938），博士研究生；王东波（ORCID：0000-0002-9894-9550），教授，博士，博士生导师；黄水清（ORCID：0000-0002-1646-9300），教授，博士，博士生导师，通讯作者，E-mail：sqhuang@njau.edu.cn。
收稿日期：2020-11-06 修回日期：2021-02-03 本文起止页码：97-104 本文责任编辑：王传清

生不同的研究结果。许多学者针对古代诗词佛经等经典著作中的异文现象展开研究。邓亚文[7]和王学军[8]分别分析了唐诗和宋词中的异文现象，总结了异文的类型、成因等内容。曾良和江可心[9]将研究重心放在了佛经异文上，基于俗字知识分析异文的成因，并认为可以通过异文来研究同源词，也可以对佛经内容进行勘误。有些学者则“专攻”某部古籍中存在的异文，江林昌[10]、周福云[11]、陈伟玲[12]分别针对《楚辞》《离骚》《怀沙》等作品，分析了异文形成的原因及其研究意义。也有学者[13-14]分别对比《隋人书出师颂》与《文选・出师颂》《太上洞渊神咒经》的《道藏》本和敦煌本，总结了其中的异文现象，并探讨了异文研究的实践意义。异文研究意义之一就是在训诂学和校勘学发挥重要作用。刘禾[15]在训诂学、校勘学意义上辨析了同书不同版本、同书不同篇、不同书载同类事以及某书引他书等情况下的异文，并总结了研究异文在训释语义、校勘古籍等方面的重要作用。边星灿[16]通过分析异文的类型、成因等内容，提出了“异文引导”这一训诂方法，以弥补异文印证的不科学性，并探讨了异文在训诂中的重要作用。王彦坤[17]、吴辛丑[18]、于亭[19]基于不同古籍文献，突出强调了异文在文字、音韵、训诂、词汇、语法及版本校勘等方面重要的学术价值。近年来在训诂和校勘方面的异文研究则细化到具体典籍中的异文，狄碧云等[20]表示异文材料是古籍校勘中的重要参考依据，并提出对《灵枢经》的异文研究将对中医学、传统语言学等方面提供帮助。薄迎迎[21]针对《楚辞・九章》中不同类型的异文进行了研究，提出这些异文对研究汉语史、楚辞训诂史以及中古时期语言文字均具有重要意义。

异文在古籍整理和源流辨别等方面的作用亦是不容忽视的。冯青[22]和陈立华[23]分别以《世说新语》和《晋书》中的异文词汇、译经《生经》中的异文现象为例分析了异文研究对《汉语大词典》编撰的作用。陈仁仁[24]从比卦异文入手，解读了今本、帛本、阜阳本和楚竹书本这4个版本的比卦异文，为《周易》版本变迁的相关研究提供线索，并提出基于异文研究还可以对比卦卦爻辞的历史信息得出融会贯通的解读。任璐[25]基于《说无垢称经》这一汉文佛典中的版本异文和同经异译本，分析了同经异译本研究、版本异文校释、同经异译对比、字际关系说明等对异文在大型语文辞书编撰起到的重要作用。章琦[26]则基于两种不同版本的唐诗《观棋》考证作者创作情况及文本文化内涵，对《观棋》的原创作者和修改作者做出判断。

也有一些研究者从与异文密切相关的诗词表意、诗情关系等角度入手对异文进行深度剖析。王骧[27]以李白《蜀道难》诗中的一处异文为研究对象，从物情事理、诗语来历、版本异同和语言表现等方面，分析了该处异文。周福云[28]列举出了《全唐诗》辑录中的十余处王维诗异文，通过比较研究的方式对异文词义与诗情关系展开分析。郭殿忱和郭志媛[29]基于《河岳英灵集》，与《唐写本唐人选唐诗》《乐府诗集》等多种集本相比较，得出李白诗异文若干，并对这些异文中的意境、声韵及诗人际遇等方面展开讨论。

除上述热点方向之外，部分学者另辟蹊径，从异文与教育教学角度进行研究。崔达送等[30]主要研究了异文与古汉语教学之间的关系，并认为异文教学有助于学生学习汉语、扩展知识面等。过雨辰[31]考证了《宿新市徐公店》的城市地点，并根据当地季节花木现象等方面，探讨版本异文出现的合理性，进而对教材编撰中可能出现异文版本的选择问题提出建议。

随着信息技术的发展，以及研究者对于古籍数字化进程的不断推动，作为古籍研究不可或缺的部分，异文研究中的异文自动发掘也逐渐成为研究热点。异文的自动发掘可以从大量的古籍文本中抽取出可用于校对的文本信息，并实现穷尽式检索[32]。这一方法有助于解决传统古籍校勘需要耗费大量人工成本的劣势，同时，经过这一过程，能够获取的信息也会更加系统全面，为有效完善古籍校勘的技术和方法提供思路。姜慧敏等[33]以《方志物产》为研究对象，提出了根据方志体例规律设计的自动合并算法，比较段落内容，求同存异，并对异文版本进行标注，实现了同地域不同时代方志版本内容的自动合并。肖磊和陈小荷[34]以《三传春秋经》为例，通过bigram计算句珠相似度匹配并去除同文的方式来实现古籍版本异文的自动发现。李越[35]从计算语言学的角度基于句子相似度算法实现了《左传》和《史记》中的同事异文自动发现，并改进了编辑距离算法，为大规模古籍异文处理提供思路。赵红[36]强调了吐鲁番文献中的异体字对汉语史语料库建设的重要意义，并认为在古汉语熟语料库中保留异文将对异文的自动检索和自动发现提供帮助。谢靖[37]基于句子匹配算法针对《黄帝内经》进行了异文自动发现研究，为中医古籍的相关研究提供了重要参考。

上述研究大多仅从汉语言文学这一视角对异文进行剖析，现有成果中针对异文自动发掘的研究还相对较少。本研究将数据源细化到经典的编年体史书《春秋》及“春秋三传”上，引入平行语料库的思想，并尝试

应用深度学习算法,为大规模语料情形下的异文自动发掘提供帮助。

3 模型应用

异文的自动发掘在某种程度上可以理解为具有语义相似性的两个句子的自动匹配,因此,根据这一特点,本研究在引入平行语料库思想的同时,尝试应用经典的支持向量机(Support Vector Machine, SVM),以及长短时记忆网络(Long Short-Term Memory, LSTM)和BERT(Bidirectional Encoder Representation from Transformers)模型中的分类任务来实现异文的自动发掘。

3.1 SVM

支持向量机被公认为是在深度学习出现之前最有效的机器学习算法之一,是一种监督学习,可以解决分类和预测等问题。它的基本思想是基于训练集在样本空间中找到一个划分超平面,将不同类别的样本分开[38]。在异文识别的任务中,对候选句对中的两句,如:"車馬曰賵"和"乘馬曰賵",分别进行表征并拼接为待分类的特征,通过大量标记候选句对训练划分"0""1"的超平面,实现异文自动发掘模型。因此,本文选用SVM作为探索异文自动发掘模型的基准,通过对比尝试获取更有效的新模型。

3.2 LSTM

长短时记忆网络[39]是循环神经网络(Recurrent Neural Network,RNN)的一种变形形式,本研究采用的是Siamese LSTM,主要用以解决二分类问题,被称之为Siamese是由于该模型左右两边的句子共享同一权重。以本研究为例,Siamese LSTM的分类过程是将两个候选句子输入到模型之中,a模型输入"克者何",b模型输入"克之者何",然后计算两者的隐层向量的曼哈顿距离(Manhattan distance)来评价句子相似度。公式如下(相似度区间为0－1)[40]:

$$D = exp\left(-\left\| h^{(left)} - h^{(right)} \right\|_1\right) \quad \text{公式(1)}$$

根据公式(1),可以确认两个句子是否为异文句对。如图1所示,其中,给定一个样本[a,b,y],x,y分别为输入和输出,y的结果为[0,1]。

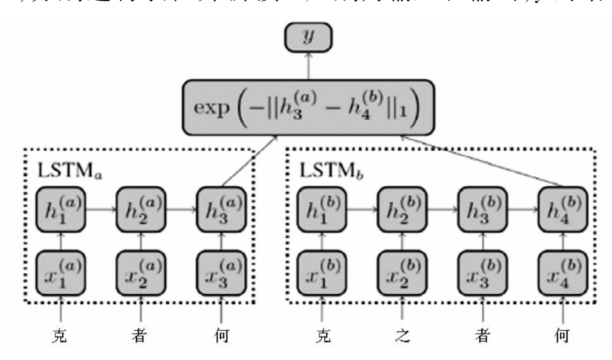


图1 Siamese LSTM模型结构

3.3 BERT

BERT是由谷歌提出的一种深度学习模型[41],即基于Transformer的双向编码器表征。其可以实现同时利用待预测词汇的上下文信息来解决需要处理的问题,进而学习句间关系,以判断两个句子是否相关。因此,BERT通常在文本分类中表现相当优异。以本实验数据中的一对异文句对为例,将"公何以不言即位"和"不书即位"两句输入到模型中,然后通过随机遮盖、替换或上下文预测的预训练方式来建立语言模型,进而预测候选句对间的关系,即预测句子类别,如图2所示:

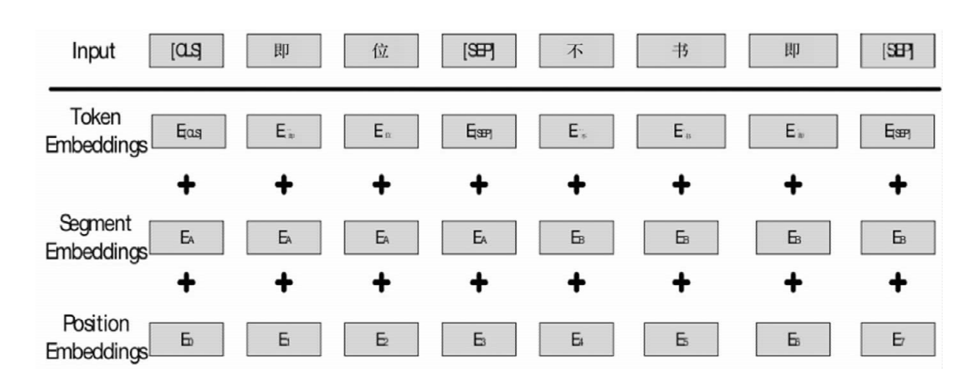


图2 BERT模型结构

4 异文标注体系构建

对于中国历代传世文献的数字化整理研究而言，“中国哲学书电子化计划”的工作首屈一指，这是一个线上电子书开放平台，其打破了纸质印刷的壁垒，收录了超过三万部著作，是目前最大的历代中文文献资料库。本研究从该平台获取了“春秋三传”，即《春秋公羊传》《春秋穀梁传》《春秋左传》的原文数据，其后对结果进行清洗、去重、校对等预处理。

4.1 数据预处理

本研究是对于同事异文自动发掘的初探，因此，优先考虑短句级别异文匹配和分类，笔者以逗号（，）句号（。）叹号（！）问号（？）分号（；）冒号（：）为句子切分符对上文数据进行分句，共获得短句 67 693 句。由于后续研究需利用每个短句的位置信息及上下文语境等，因此，对每个短句赋予标号（见表 1），其中序号为顺序标号，处理后结果见图 3。

表 1 短句标号规范

短句所属典籍	标号方式
《春秋》	0－序号
《春秋公羊传》	1－序号
《春秋穀梁传》	2－序号
《春秋左传》	3－序号
《春秋左传附》	4－序号
其他（包括篇名、卷名等）	9－序号

1-131眛者何
1-132地期也
1-133
9-
2-134穀
2-135及者何
2-136內爲志焉爾
2-137儀、字也
2-138父猶傅也
2-139男子之美稱也
2-140其不言邾子何也
2-141邾之上古微
2-142未爵命於周也
2-143不日、其盟渝也
2-144眛、地名也
2-145
9-
3-146左
3-147三月
3-148公及邾儀父盟于蔑
3-149邾子克也
3-150未王命
3-151故不書爵
3-152曰儀父、貴之也
4-7生隱公
4-8宋武公生仲子
4-9仲子生而有文在其手
4-10曰爲魯夫人
4-11故仲子歸于我
4-12生桓公而惠公薨
4-13是以隱公立而奉之
4-14
0-1
0-15元年
0-16春
0-17王正月
0-18
9-
1-19公
1-20元年者何
1-21君之始年也
1-22春者何
1-23歲之始也
1-24王者孰謂
1-25謂文王也
1-26曷爲先言王而後言正月
1-27王正月也
1-28何言乎王正月

图 3 分句示例

4.2 异文标注规范

根据异文定义对经过预处理的“春秋三传”及《春秋》原文进行异文标注，异文句一一对应，具体规范如下：①语义完全对应且文本相似度较高。例如“成公意也”和“成公志也”表意相同，标注为异文句。②部分语义及文本对应。以短句为单位匹配，如“公將平國而反之桓”和“將以讓桓也”对应形成异文句对。③语义相似度高而文本相似度较低，但包含同义词等。同样作为异文句对处理，如“盟納季子也”和“請復季友也”，都表达了此次会盟的目的是请季友回国，但表述差异相对较大。④文本几乎无相似之处，但表达同一事件。如“則齊國盡子之有也”和“舉國而授”作为异文句对标注。⑤部分成分存在省略或简写。例如“衞孫良夫率師侵宋”和“晉伯宗、夏陽說、衞孫良夫、甯相、鄭人、伊雒之戎、陸渾、蠻氏侵宋”这两句其实表达的均是侵宋这一事件，但后者对于发动战争一方有更为详尽的表述，这两个短句也作为异文句对处理。再如，“諸侯救鄭”相较于“公會齊人、宋人、邾婁人救鄭”运用“诸侯”来代指参与救郑的诸侯国。⑥时间及表时间的季节等，均不作为异文句对标注。如“元年”“春”“三月”等。⑦文本相似度高，但语义表达核心不同，不作为异文句对。如“宋師及齊師戰于甗”表达宋齐双方在甗交战，而“宋敗齊師于甗”则体现了战败方即战果，因此这两句不作为异文句对处理。⑧并未包含于上述 7 项规范中的情况，笔者均会查阅相关资料，具体语句具体分析，并进行统一，确保标注的规范性。

笔者根据上述规范进行人工标注后，共获得异文句对 1 692 对。

4.3 异文平行语料库构建

候选句对集以异文句子对和非异文句子对 1∶1 的比例构成，获得候选句对 3 384 对，其中异文句对根据 4.2 节的规范经人工标注生成，非异文句对中各句子均来自 4.1 节的数据，句 b 由其他两传中句 a 所在段落对应的段落中自动生成，通常为异文句对中句 a 对应句 b 后的第 5 句。之后，借助平行语料库的形式，将句子对一一对应，采用“0－1”分类的方式，0 表示非异文句子对，1 表示异文句子对，异文平行语料库部分数据见表 2。

5 异文自动发掘模型实验结果与评价

5.1 实验环境与参数设置

本实验采用的操作系统为 ubuntu 16.04，内存为 16GB DDR4，显存为 4GBGDDR5，CPU 为 Intel（R） Core

表 2　候选句对示例

标注	句 a 序号	句 b 序号	句 a 文本	句 b 文本
1	0－17	3－101	王正月	王周正月
1	1－31	2－61	成公意也	成公志也
0	1－2141	3－2264	假甚也	羽父先會齊侯、鄭伯伐宋
0	1－2161	2－2300	俠者何	此其日何也
1	1－33	2－65	公將平國而反之桓	將以讓桓也
1	2－142	3－150	未爵命於周也	未王命
1	1－168	2－188	克之者何	克者何
0	1－2594	2－2668	隱何以無正月	繼故而言卽位則是與聞乎弑也
1	1－169	2－191	殺之也	能殺也
0	1－3388	3－3426	己丑之日死而得	齊侯、鄭伯朝于紀

(TM) i5－4590 CPU @ 3.30GHz,GPU 型号为 NVIDIA Quadro K1200。

为确保 3 种模型的实验结果具有可比性,输入数据均为经过统一处理的相同语料,并在相同实验环境下进行。3 种模型参数设置见表 3。

5.2　评价指标

此外,对模型性能的评价,本研究采用的指标为准确率(Precision)、召回率(Recall)、F 值(F-Measure)。具体计算公式如下:

表 3　模型参数设置

模型	参数设置
SVM	惩罚系数 2.0，核函数类型为 RBF，核函数系数为 0.5，最大迭代次数 10000，缓冲大小为 4000，随机种子大小 42
LSTM	预训练字向量维度 128 维,神经元数量 200,每批数据量大小 200,最大训练时期数 100
BERT	预训练模型:BERT-Base，Chinese：Chinese Simplified and Traditional，12－layer，768－hidden，12－heads，110M parameters,该模型共 12 层,隐层 768 维,12 头模式,110M 个参数。最大截断长度 256,训练批次大小 32,学习率 2e－5,迭代次数 10.0 次

$$\text{准确率 } P = \frac{\text{正确识别的句对}}{\text{正确识别的句对} + \text{被错误识别的句对}} \times 100\% \quad \text{公式(2)}$$

$$\text{召回率 } R = \frac{\text{正确识别的句对}}{\text{正确识别的句对} + \text{未被识别的句对}} \times 100\% \quad \text{公式(3)}$$

$$\text{调和平均值 } F = \frac{2 \times P \times R}{P + R} \times 100\% \quad \text{公式(4)}$$

5.3　实验结果

本实验分别运用 SVM、LSTM 和 BERT 模型在同一数据集上进行十折交叉验证,评价指标数值如表 4 所示:

表 4　十折交叉实验评价指标结果

	SVM			LSTM			BERT		
	P/%	R/%	F/%	P/%	R/%	F/%	P/%	R/%	F/%
1	22.79	45.99	30.48	53.44	52.10	52.76	54.92	54.09	54.50
2	19.44	42.39	26.66	44.55	46.65	45.58	56.96	54.96	55.94
3	21.10	44.21	28.57	56.49	53.96	55.20	56.45	54.85	55.64
4	26.91	45.70	31.56	48.81	49.28	49.04	57.08	55.13	56.09
5	21.61	45.40	29.28	52.40	51.47	51.93	61.25	57.36	59.24
6	22.35	45.10	29.88	50.28	50.17	50.22	54.99	54.07	54.53
7	21.42	43.92	28.80	49.29	49.57	49.43	60.68	58.33	59.48
8	21.34	43.62	28.66	48.75	49.24	49.00	52.27	51.69	51.98
9	27.77	43.92	30.26	48.33	48.98	48.65	62.33	60.37	61.34
10	24.25	46.69	31.92	49.53	49.71	49.62	61.36	59.20	60.26
均值	22.90	44.69	29.61	50.19	50.11	50.14	57.83	56.00	56.90

通过表 4 可以看出,整体上 LSTM(F = 50.14%)和 BERT(F = 56.90%)的效果要明显优于 SVM(F = 29.61%),SVM 的 F 值甚至均低于 30%。虽然 LSTM 和 BERT 两者的 F 值差值相对较小,但后者也有一定幅度的提升,BERT 的 3 项评价指标几乎均高于 LSTM 对应实验组别的评价指标,并且,BERT 最优模型的 F 值达到了 61.34%。虽然得出的评价指标数值似乎并没有非常高,但是与前人的实验结果相比较也是相对理想的,特别是,在此之前运用深度学习甚至是机器学习进行异文自动发掘的研究均非常少,李越[35]改进编辑距离算法实现同事异文自动发现最好结果为 P = 46.02%,R = 90.15%,F = 60.93%,本实验最优模型 F 值高于已有算法结果,且准确率有明显提升。

本实验将深度学习模型应用到“春秋三传”异文自动发掘中,并与经典机器学习算法 SVM 在相同数据集上的实验结果进行对比,证明深度学习在该研究领域中的表现有明显优势,因此,在异文自动发掘模型的选择上,BERT 的表现较为出色,而 LSTM 也相对较好,

两者都明显优于较为经典的 SVM 模型。

6 基于异文自动发掘结果分析

本文对标注得到的 1 692 对异文句对进一步分析,异文句对的数量分布如表 5 所示,《春秋》《公羊传》21 对句对互为异文,《春秋》《穀梁传》19 对句对互为异文,《春秋》《左传》970 对句对互为异文,《公羊传》《穀梁传》中存在 513 对异文,《公羊传》《左传》共有 89 对异文句对,《穀梁传》《左传》共有 80 对异文句对。

表 5 异文句对分布情况(单位:对)

典籍来源	春秋	公羊传	穀梁传	左传
春秋	–	20	19	969
公羊传	1	–	502	88
穀梁传	0	11	–	78
左传	1	1	2	–

"春秋三传"有各自的语言风格,《左传》大部分行文文辞简洁,通常运用和《春秋》相同或相似的表述一笔带过、简明扼要,但对于需要特别扩展的事件则不厌其详,能够清晰地叙述事件来龙去脉的同时,又使事件主人公的形象跃然纸上。《左传》中描绘生动的人物对话也是其语言特色之一,这两点也同时为异文自动识别增大了难度,但这也是异文自动发掘研究的意义之一。除此之外,《左传》也将不少笔墨放在了对于事件发生这一时期的背景上,多以"左附"的方式出现,本研究中"左附"的内容并没有纳入异文标注体系之中,主要是由于其一般为介绍某一时期或事件的前因后果、环境背景等,因此很少与其他 3 部典籍中的文本形成异文。

相对而言,《穀梁传》和《公羊传》的叙事风格则更为相似,大多采用设问句这种一问一答的方式行文,由上一个问题的答案引出下一个问题,逻辑清晰、条分缕析,内容多集中于记录某一事件的原因,该事件具有哪些特殊性或者与其他事件有哪些关联或共同点等,这些异文句对识别难度相对较小。

但这两本典籍中有时也会出现对于同一事件有着不同甚至相反的解释。

例:

《春秋》 "齊人伐山戎。"

《公羊传·庄公三十年》 "此齊侯也,其稱人何?貶。"

《穀梁传·庄公三十年》 "齊人者、齊侯也。其曰人何也?愛齊侯乎山戎也"

《公羊传》在解释《春秋》称齐侯为"齐人"的原因时,对齐侯表达了贬抑之情,而《穀梁传》则解释为对齐侯的爱惜,不想齐侯与山戎并称。这种情况由于候选句对表意完全相反,因此,在本研究中不作为异文句对进行处理。

7 结语

数字人文是新技术与人文研究有机结合的良好方式。近年来,随着古籍数字化进程的不断推进,古籍相关研究在数字人文领域逐渐占有一席之地,而异文则是古籍研究,特别是古籍数字化不可缺少的重要部分。异文的自动发掘不但可以大幅度降低传统古籍校勘的人工成本,同时也为新技术与典籍研究的融合验证了可行性。因此,本研究以《春秋》及"春秋三传"为实验语料,引入常用于文本翻译领域的平行语料库思想,构建了异文平行语料库及标注规范,并通过对 SVM、LSTM 和 BERT 模型实验结果的对比分析,发现深度学习模型在异文自动发掘方面有很好的表现,这说明深度学习和平行语料库在异文相关研究中可以发挥较大的作用,为异文自动发掘提供可行性方案,值得扩展和复用。笔者将在后续的研究中不断探索新技术,提高异文自动发掘模型性能的同时,也会将应用语料延伸到更多典籍之中。

此外,通过对识别出的异文句对进行分析,本研究可从以下几个方面改进和探索:①对于低频异文类型的识别,尝试改进模型或算法,比如文本相似度很高的非异文句对以及文本相似度很低的异文句对识别难度较大,常出现识别错误或无法识别的现象。②本研究目前只尝试了短句一对一匹配的形式,采用将多对一的句子转化为一对一的形式,即忽略部分短句,只将语义核心短句标注为异文句对,这可能也会对异文自动发掘效果产生影响,后续研究将尝试更多类型的异文句对标注形式。③"春秋三传"均属于编年体史书,本研究从该类型史书入手,未来的研究将扩展到纪传体史书例如《史记》等典籍上,也希望后续可以应用到更多类型的古籍异文自动发掘中。④本研究采用的模型均未加入上下文语境特征、同义词词典等,后续笔者将利用序号自动获取异文句对的上下文,辅助提高模型识别效果,同时引入人名、官职名、各词性同义词词典等,以求获得更具优势的异文自动

发掘方法。

参考文献：

[1] 黄沛荣. 古籍异文论析[J]. 汉学研究，1991，9（2）:395.
[2] 李娟.《史记》《汉书》异文中的同源词研究[J]. 湖北师范学院学报(哲学社会科学版),2011,31(4):60－63.
[3] 李娟.《史记》《汉书》异文的训诂价值研究[D]. 黄石:湖北师范学院,2012.
[4] 罗积勇. 异文与释义[J]. 古籍整理研究学刊，1986(2):58－60.
[5] 王彦坤. 试论古书异文产生的原因[J]. 暨南学报:哲学社会科学版，1989(4):78－85.
[6] 石云孙. 话语中的异文[J]. 安庆师范学院学报(社会科学版)，1996(2):2－8.
[7] 邓亚文. 论唐诗异文[J]. 湖北科技学院学报，2002，22(5):68－70.
[8] 王学军. 宋词异文探微[J]. 文教资料，2010(18):32－36.
[9] 曾良，江可心. 佛经异文与词语考索[J]. 古汉语研究，2013(2):43－48.
[10] 江林昌.《楚辞》异文考例[J]. 文献，1991(3):3－14.
[11] 周福云.《离骚》异文例释[J]. 淮阴师范学院学报(哲学社会科学版)，1993(2):20－24.
[12] 陈伟玲.《怀沙》异文考辨[J]. 职大学报，2007(1):46－47.
[13] 易敏.《隋人书出师颂》及《文选》异文[J]. 井冈山师范学院学报，2005(1):5－8.
[14] 牛尚鹏.《太上洞渊神咒经》异文考辨[J]. 长江师范学院学报，2016,32(1):73－78.
[15] 刘禾. 异文与训校[J]. 东北师大学报(哲学)，1986(2):62－69.
[16] 边星灿. 论异文在训诂中的作用[J]. 浙江大学学报(人文社会科学版)，1998(3):135－140.
[17] 王彦坤. 略论古书异文的应用[J]. 暨南学报:哲学社会科学版，1987(1):75－81.
[18] 吴辛丑. 简帛异文的类型及其价值[J]. 华南师范大学学报(社会科学版)，2000(4):37－42.
[19] 于亭. 异文用于训诂实践的历史透视[J]. 长江学术，2009(3):131－138.
[20] 狄碧云，孙兆杰，范登脉. 浅谈《灵枢经》的异文研究[J]. 中医文献杂志，2013，31(3):12－14.
[21] 薄迎迎.《楚辞疏・九章》异文研究[J]. 语文学刊，2016(24):59,107.
[22] 冯青. 异文词汇与词汇史研究[J]. 哈尔滨师范大学社会科学学报，2010,1(1):52－55.
[23] 陈立华.《生经》异文研究[D]. 长沙:湖南师范大学，2011.
[24] 陈仁仁. 比卦异文解读[J]. 中国哲学史，2010(3):54－62.
[25] 任璐.《说无垢称经》异文研究[D]. 贵阳:贵州师范大学，2015.
[26] 章琦.《观棋》作者、异文考[J]. 北京社会科学，2016(2):83－89.
[27] 王骧. 李白《蜀道难》诗的一处异文[J]. 高校教育管理，1990(2):16－17.
[28] 周福云. 王维诗异文探索[J]. 台州学院学报，1998(1):75－77.
[29] 郭殿忱，郭志媛. 李白诗异文考——以《河岳英灵集》为中心[J]. 绵阳师范学院学报，2014，33(1):12－18.
[30] 崔达送，詹绪左，储泰松，等. 异文比较与古汉语教学[J]. 滁州学院学报，2008，10(1):22－25.
[31] 过雨辰.《宿新市徐公店》异文考[J]. 剑南文学(下半月)，2016(5):33－34.
[32] 鞠明库. 古籍数字化与传统文献学[J]. 清华大学学报(哲学社会科学版),2011,26(5):154－158,161.
[33] 姜慧敏，白振田，周金水，等. 同一地域不同时代方志版本内容自动合并的研究与实现[J]. 广西地方志，2008(5):29－32.
[34] 肖磊，陈小荷. 古籍版本异文的自动发现[J]. 中文信息学报，2010，24(5):50－55.
[35] 李越.《左传》《史记》同事异文自动发现及分析[D]. 南京:南京师范大学，2014.
[36] 赵红. 吐鲁番文献与汉语语料库建设的若干思考[J]. 南京师范大学文学院学报，2014(3):155－158.
[37] 谢靖. 基于句子匹配的《黄帝内经》异文自动发现研究[J]. 科技视界，2015(35):53－54.
[38] HEARST M A，DUMAIS S T，OSUNA E，et al. Support vector machines[J]. IEEE intelligent systems & their applications，1998，13(4):18－28.
[39] GRAVES A，SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural networks，2005，18(5/6): 602－610.
[40] MUELLER J,THYAGARAJAN A. Siamese recurrent architectures for learning sentence similarity[C]//Proceedings of the thirtieth AAAI conference on artificial intelligence. Phoenix,Arizona:AAAI，2016:2786－2792.
[41] DEVLIN J，CHANG M W，LEE K，et al. Bert: Pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 annual conference of the North American chapter of the Association for Computational Linguistics. Minneapolis: ACL,2019:4171－4186.

作者贡献说明：

梁媛:进行数据处理,起草论文;
王东波:提供研究思路，设计研究方案;
黄水清:提出相关概念及整体研究思路,修订完稿。

Research on Automatic Mining of Variants Expressing the Same Event in the Ancient Books

Liang Yuan[1,2] Wang Dongbo[1,2] Huang Shuiqing[1,2]
[1] College of Information Management, Nanjing Agricultural University, Nanjing 210095
[2] Research Center for Humanities and Social Computing, Nanjing Agricultural University, Nanjing 210095

Abstract: [Purpose/significance] Variations are a common phenomenon and also an important research object in ancient books. The traditional collation of ancient books is to manually search for materials, including variations from a large number of ancient books. This work is not only time-consuming, laborious, and heavy, but the data may not be accurate and comprehensive. Automatic mining of variant sentences through computers can obtain effective information from larger-scale corpus. In addition, the collation method combined with automatic mining of variant sentences can realize exhaustive retrieval, which is of great significance to the collation of ancient books. It provides new ideas and methods for the collation research of ancient books in the new period. [Method/process] This research automatically mined the variant sentences in Three Biographies of the Spring and Autumn Period, combining deep learning and introducing parallel corpus commonly used in the field of machine translation. Subsequently, this study compared LSTM and BERT models' results with the classic SVM model and further explored and analyzed the related content of the variants expressing the same event with different descriptions in two ancient books. [Result/conclusion] The experiment obtained a deep learning model for automatic mining of variants expressing the same event suitable for Three Biographies of the Spring and Autumn Period. It proves the feasibility of integrating new technologies such as deep learning into the construction of ancient books' knowledge base. Meanwhile, the combination of deep learning and parallel corpus can play a more significant role in studying variant sentences and provide practical support for applying digital humanities in the Chinese language and literature.

Keywords: Three Biographies of the Spring and Autumn Period variants BERT automatic mining digital humanities